מאגר הפומות בעברית – פרויקט שתחילתו במאה מיליון מילים ותכליתו 31 פומות

ורד זילבר-ורוד, משה לטין ועמי מויאל
המרכז לעיבוד שפה, אפקה – המכללה האקדמית להנדסה בתל-אביב

החוג הישראלי לבלשנות ע"ש חיים רוזן
המפגש העשרים ותשעה
4 בפברואר 2013

ACLP
Afeka Center for Language Processing

AFEKA TEL-AVIV ACADEMIC COLLEGE OF ENGINEERING

מטרת המחקר

- מידול אקוסטי של 31 הפונמות בעברית המדוברת בישראל.
- תשתית חיונית לכל טכנולוגיית דיבור, בראש ובראשונה טכנולוגיית זיהוי דיבור (Automatic Speech Recognition).

תנועות

i

e

a

o

u

עיצורים	Bilabial	Labial-dental	Alveolar	Post-alveolar	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b		t d			k g			ʔ
Nasal	m		n						
Fricative		f v	s z	ʃ ʒ		x	χ ʁ	ʕ	h
Approximant	w				y				
Lateral approximant			l						

עיצורים מחוככים: /tʃ/ /dʒ/ /ts/

דרישות:

ייצוג הפונמות וסביבתן בכמות הופעות (occurrences) מספקת לכל פונמה ומספיק חזרות של כל צמד פונמות (diphone) ורצף של 3 פונמות (triphone) שאופייניים לשפה.

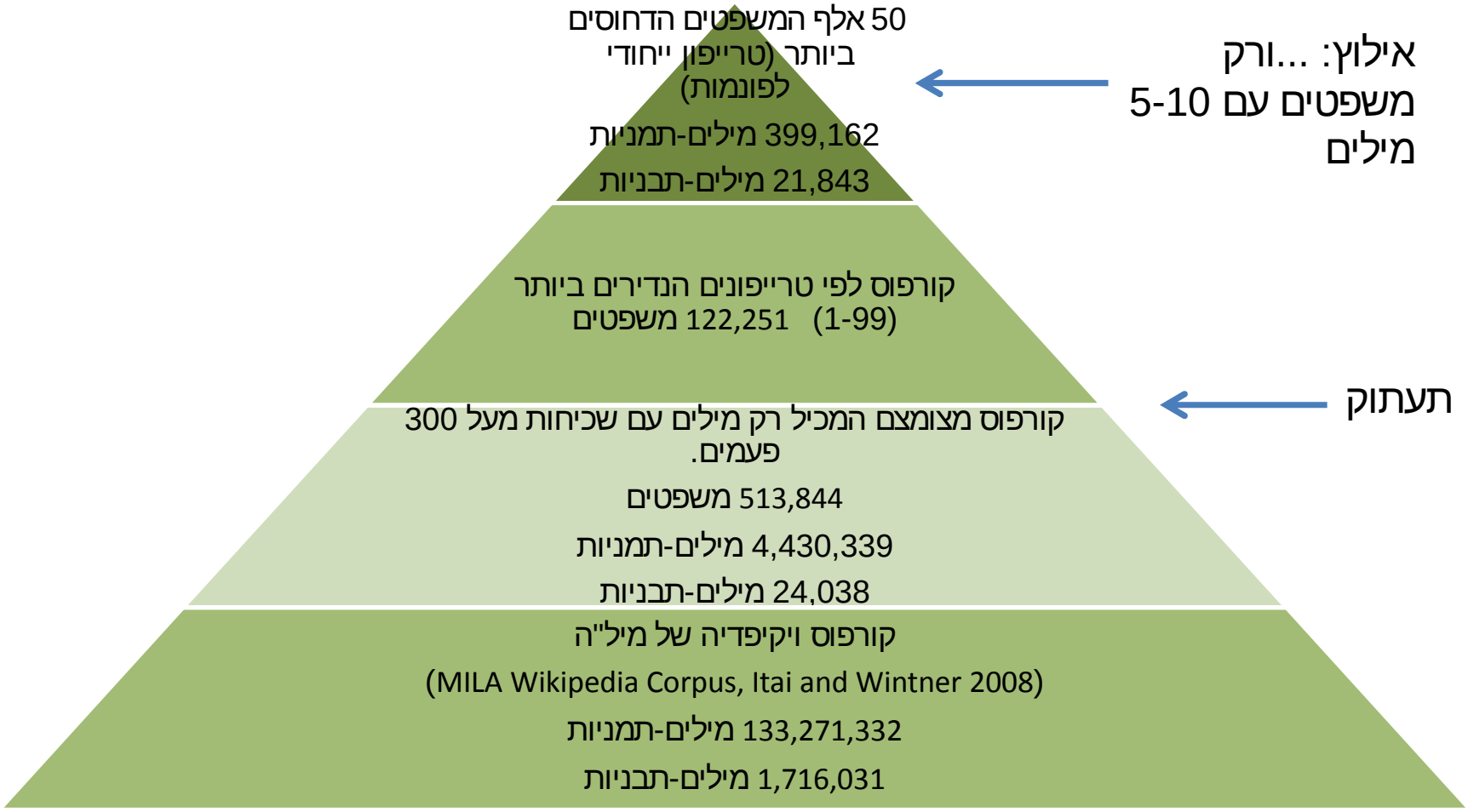
"It is crucial that all phonemes are represented in the speech corpus **in sufficiently high numbers**." (Gibbon, Moore, and Winski (eds.), 1997)

כיצד ניתן להשיג ייצוג של פונמות?

- ניסוח משפטים.
- קורפוס מאוזן פונטית (phonetically balanced)
  - ייצוג יחסי, לפי השכיחות בשפה
- קורפוס עשיר פונטית (phonetically rich)
  - ייצוג כמעט זהה לכל הפונמות

השלבים הנדרשים לכינון קורפוס "עשיר פונטית":

1. בחירת קורפוס
2. המרת אותיות לפוננמות (G2P) = תעתוק.
3. בחירת משפטים שמכילים את המספר הרב ביותר של פוננמות-TYPES, לא של פוננמות-TOKENS (כלומר לא לפי כמות הפוננמות אלא לפי עושר פונמיאלי).
4. בחירת משפטים שבהם יש את מספר ה פוננמות- TYPES הרב ביותר שאינן מופיעות במשפט הראשון, וכן הלאה. סיום התהליך לאחר שיש ייצוג מלא של כל הפוננמות בשפה.
5. כל התהליך הזה חוזר על עצמו כדי לקבל ייצוג של פוננמות מספיק פעמים (פרמטר כמותי).


50 אלף המשפטים הדחוסים ביותר (טרייפון ייחודי לפוננמות)
399,162 מילים-תמניות
21,843 מילים-תבניות
קורפוס לפי טרייפונים הנדירים ביותר (1-99) 122,251 משפטים
קורפוס מצומצם המכיל רק מילים עם שכיחות מעל 300 פעמים.
513,844 משפטים
4,430,339 מילים-תמניות
24,038 מילים-תבניות
קורפוס ויקיפדיה של מיל"ה
(MILA Wikipedia Corpus, Itai and Wintner 2008)
133,271,332 מילים-תמניות
1,716,031 מילים-תבניות
אילוץ: ...ורק משפטים עם 5-10 מילים
תעתוק

שכיחות פונמות וצירופי פונמות (די, טרי)

	מקסימום מונו	מינימום מונו	מקסימום די	מינימום די	מקסימום טרי	מינימום טרי
קורפוס 50 אלף משפטים	a (344,100)	ʒ (279)	h a (55,250)	ʒ p 1 "הטקסט במקורו מהערך ז'ורז' פרק"	t h a (10,641)	z v k 1 "מצודת ציון (שיר השירים ז ו) כמין צבע אדומה."

שכיחות פונמות וצירופי פונמות (די, טרי): שפה כתובה לעומת דיבור ספונטני

	מקסימום מונו	מינימום מונו	מקסימום די	מינימום די	מקסימום טרי	מינימום טרי
קורפוס 50 אלף משפטים (ראו נספח)	a (344,100)	ʒ (279)	h a (55,250)	ʒ p (1)	t h a (11,176)	z v k (1)
קורפוס ספונטני COSIH (זילבר-ורוד 2011) (ראו נספח. כולל היסוסים)	A (20,553)	ʒ (7)	h a / ma (2,000+)	z s (1)	a n i (934)	z z o (1)

- השוואה לפרויקט TIMIT (60 פונמות)

Sentence Type	Sentences	Speakers	Total	Sentences/ Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342	630	6300	10

- הרצוי: ייצוג של כל הרצפים האפשריים בשפה
  - הרצף [p d i] אינו מופיע במשפטים שנאספו על אף שכן נמצאים הרצפים [pda pde pdo pdu].

- המצוי: ייצוג הרצפים שמוצו מקורפוס של 100 מיליון מילים
  - "כדי להשלים את המחקר (על חוקי צירופי הפוננמות, חוקים פונוטקטיים) יש להכין מאגר גדול של עברית בתעתיק פונטי ופוננמי ולגלות על פיו לא רק את הצירופים האפשריים, אלא גם את תפוצתם. מידע כזה הוא סעיף בדקדוק לשוננו, ולכן עלינו לחשוף אותו..." (לאופר תשנ"א, 12)

חוקי צירופי הפונמות, חוקי פונוטקטיקה, הם תלויי שפה

אילוצים:

- מבנה ההברה בעברית (CCC)V(CCC) (לאופר תשנ"א)
- סולם הצליליות (SSP – Sonority Sequencing Principle)
- מקום החיתוך, אופן החיתוך
- [pli/ a] 'פליאה' אבל [stira] 'סטירה'!

Sonority scale	חצי תנועה	שוטפים	אפיים	חוככים	סותמים
חצי תנועה			a		
שוטפים		l			
אפיים			n		
חוככים					
סותמים	p				t

מטריצת הפונמות: רצפי עיצורים בשפה (CCVCC)

		stress	unstress	stress	unstress
		p		**b**	
onset (XY)	**p**	"**p**etel	**p**o"lin	bpVCC	bpVCC
onset (YX)		Rits"**p**a	ipa”ron	pbVCC	pbVCC
coda (XY)				CCVbp	CCVbp
coda (YX)				CCVpb	CCVpb
onset (XY)	**b**	pbVCC	pbVCC	"beReX	be"ReX
onset (YX)		bpVCC	bpVCC	niS"ba	"?aba
coda (XY)		CCVpb	CCVpb	...	...
coda (YX)		CCVbp	CCVbp	...	...
		s		...	...
onset (XY)	**f**	fsVCC	fsVCC	...	...
onset (YX)		"sfinks	sfaRa"di sfi"na sfat ha"jam	...	...
coda (XY)		CCVfs	CCVfs	...	...
coda (YX)		CCVsf	CCVsf	...	...
onset (XY)	**v**	vsVCC	vsVCC	...	...
onset (YX)		svVCC	svi"va	...	...
coda (XY)		CCVvs	CCVvs	...	...
coda (YX)		CCVsv	CCVsv	...	...

מטריצת הפונמות: רצפי עיצורים אפשריים בשפה לעומת מצאי בקורפוס

רצפים אפשריים בגבולות ההברה	a	e	i	o	u
p d **V**	+	+	-	+	+
p g **V**	+	-	+	-	-
p n **V**	+	+	+	+	-
p z **V**	+	+	-	-	-
p S **V**	+	+	+	-	+
p X **V**	+	-	-	-	-
p ts **V**	+	-	+	-	+
b X **V**	+	+	+	-	-
t X **V**	+	+	+	-	+
V j t	+	+	-	+	+
d X **V**	+	-	+	-	-
D XT **V**	+	+	+	+	+
k z **V**	+	+	+	+	-
k X **V**	+	-	-	-	-
g X **V**	-	-	-	-	-

הניתוח ההשוואתי חושף כמה היבטים:

- השפעת הבחירה המוקדמת בסט הפוננמות ובחוקי התעתוק לפיהם תועתקו המשפטים.
  - למשל, הכללת הסותם הסדקי [?] בסט הפוננמות, או קביעה כי הוא יתועתק רק בתחילת הברה מוטעמת, בהכרח משפיעים על תוצאות הרצפים שהתקבלו. לדוגמה הרצף [vue] במשפט "בחודשי המרידה עז**בו א**ת הונגריה כרבע מיליון פליטים הונגרים".
- השפה המדוברת מאפשרת רצפים שכביכול אינם אפשריים ברמת המילה,
  - כגון [wwo] ברצף המילים "וואו **וו**אוו".
  - או צירוף הפוננמות [rfg] שהתקבל פעם אחת בקורפוס, עקב תעתוק של המשפט: "שׂ**רְף ג**ם שימש כמאמן הנבחרת הלאומית של ישראל בכדורסל גברים."
  - בכך המחקר תורם לחקר הפונולוגיה בעברית המדוברת בהתייחס לרמות הפרוזודיות השונות, החל ברמת המילה (הפרוזודית) וכלה ביחידת האינטונציה.
- לעומת זאת, בתהליך מיצוי המשפטים "אבדו" רצפים אפשריים בשפה.

המשך הפרויקט

- שילוב רצפים אפשריים בשפה שאינם מופיעים ב-50 אלף המשפטים להקלטה.
- הקלטות ותעתוקים המשקפים את אופן ההגייה בפועל.

todaRaba

נספח: מצאי פונמות בקורפוס דיבור ספונטני

Total No. of Word Types: 28
Total No. of Word Tokens: 117,536

1	20553	a
2	14669	e
3	8988	i
4	7270	m
5	7100	l
6	6897	o
7	6221	t
8	4643	n
9	4331	R
10	4165	h
11	3997	S
12	3666	k
13	3548	X
14	3252	j
15	2748	v
16	2607	u
17	2505	z
18	2474	b
19	2075	d
20	1724	s
21	1210	ts
22	989	g
23	960	f
24	772	p
25	78	w
26	49	tS
27	38	dZ
28	7	Z

Total No. of Word Types: 28
Total No. of Word Tokens: 21,843,070

1	3898936	a
2	2458405	e
3	1625533	i
4	1374107	m
5	1239149	l
6	1192523	t
7	1122641	o
8	976604	R
9	877492	h
10	817748	n
11	744567	X
12	735492	S
13	664370	k
14	626049	j
15	584213	u
16	523874	b
17	498836	v
18	462244	d
19	337849	s
20	236129	z
21	234633	ts
22	230061	f
23	175947	g
24	150020	p
25	31042	dZ
26	21571	w
27	2762	tS
28	273	Z

נספח: מצאי פונמות בקורפוס וויקיפדיה 2010 (מילים בשכיחות מעל 300)